Anticipating discussion activity on community forums

Anticipating Discussion Activity on Community Forums

Matthew Rowe, Sofia Angeletou and Harith Alani
Knowledge Media Institute
The Open University
Milton Keynes, UK
Email: {m.c.rowe, s.angeletou, h.alani}@open.ac.uk

***Abstract*—Attention economics is a vital component of the Social Web, where the sheer magnitude and rate at which social data is published forces web users to decide on what content to focus their attention on. By predicting popular posts on the Social Web, that contain lengthy discussions and debates, analysts can focus their attention more effectively on content that is deemed more influential. In this paper we present a two-step approach to anticipate discussions in community forums by a) identifying seed posts - i.e., posts that generate discussions, and b) predicting the length of these discussions. We explore the effectiveness of a range of features in anticipating discussions such as user and content features, and present *focus* features that capture the topical concentration of a user. For identifying seed posts we show that content features are better predictors than user features, while achieving an F_1 value of 0.792 when using all features. For predicting discussion activity we find a positive correlation between the focus of the user and discussion volumes, and achieve an nDCG@1 value of 0.89 when predicting using user features.**

***Index Terms*—Social Web, Communities, Prediction, Discussions**

I. Introduction

Many social media systems have quickly become the *de facto* forums where web users ask questions, publish their status and discuss their opinions about events, policies and a plethora of current issues. Analysing content on social media is therefore rapidly growing into an attractive business for real-time monitoring of brands, public opinions and markets. This type of business intelligence tends to be relatively quick and low cost given the sheer amount of shared content that is constantly updated and easily accessible. According to Forrester, it is estimated that the software market for social network analytics in the US alone will reach $1 billion in 2014 [1].

This unprecedented, rapid growth of content is naturally producing new challenges for its analysis and use. One such challenge often faced by analysts is how to identify which particular content is likely to generate more attention from the community than others. Predicting the popularity of posts helps focussing the attention of human and computerised analysts and information managers more quickly and efficiently to content that is deemed auspicious and influential.

To make such predictions accurately, we need to understand the impact that the various community and content features have on how much attention is generated in the community for particular users and posts. An understanding of such factors would in turn allow content creators to shape their content into a more suitable and receivable form, thus maximising the likelihood of instigating a lengthy discussion. In our previous work [2] we predicted discussions on the Microblogging platform Twitter, and found that certain features were associated with increased discussion activity - i.e., the greater the broadcast spectrum of the user, characterised by in-degree and list-degree levels, the greater the discussion activity. In this paper we extend our analysis to discussion forums, exploring two research questions: a) *which features are key for stimulating discussions?*, and; b) *how do these features influence discussion length?*

In exploring the above questions we present a two-stage approach to predict discussion activity levels on community forums that *first*; identifies seed posts - i.e., thread starters that yield at least one reply, and *second*; predicts the discussion activity that such seed posts will generate. Through experiments over one year's worth of data from the Irish community message board Boards.ie,[1] we were able to assess our findings and thus discover the prevalent features that are associated with discussions on community forums.

A. Main Contributions

The main contributions in this paper are fourfold, and are listed as follows:

1) Identification of seed posts through a classification task. Testing a range of features of posts and their authors when combined with different classification models. We achieve an optimum F_1 level of 0.792 using all features and the J48 classifier, and identify a post's content as being key in starting a discussion.
2) *Focus* features to capture the topical concentration of users. In particular we introduce the notion of '*Forum Likelihood*', designed to combine past topic history of a user with the topic of their new post in the form of a likelihood estimate.
3) Prediction of discussion activity levels using various regression models combined with different features. We observe the difference in predictive performance that certain features hold and find that the content of the

[1] http://www.boards.ie/

post combined with the focus information of the user yields consistently accurate predictions.

4) Contrast our findings against our previous work on predicting discussions on Twitter [2]. We show that discussions in these communities are driven by different key factors, and that unlike Twitter, the reputation and role of the user within the community has less effect.

In the next section we cover related work. Section III describes the community features we analyse. Our dataset is detailed in section IV. Section V presents our classification approach and experiment to identify forum posts that yield replies. Section VI describes a regression model and experiment on predicting discussion levels in Boards.ie, followed by a discussion in section VII and onclusions in section VIII.

II. Related Work

Measuring and predicting popularity of users and content has been the focus of much research in the past few years. In terms of content, it has been found that the number of early comments and their quality and characteristics can be used to predict ranking of stories on Digg [3], content popularity on Digg and YouTube [4], best replies in Yahoo! Answers [5], [6], and story popularity on Slashdot [7]. Sentiment and polarity of content have also been successfully used to predict ratings of comments in YouTube [8].

In terms of user features, Hsu et.al [9] extended the ranking prediction of Digg stories by mixing content features (e.g., informativeness and complexity) with user features (e.g., reputation, number of friendships). Cha et. al [10] measured user influence in Twitter from their number of followers, retweets and mentions. To measure attention generation in Twitter, Suh et. al [11] analysed content and user features, and found that the existence of hashtags or URLs in posts does not influence attention generation (i.e., retweetability) whereas the number of followers and followed users does. Similarly, Hong et. al [12] predicted whether a post will be retweeted and how much retweet volume it will generate based on user (network statistics) and content features (TF-IDF and posting time) as well as post topics.

Our previous work on measuring and predicting response generation in Twitter [2] extended the approaches above by not only analysing a richer mix of content and user features, but also for introducing a two-stage model for predicting which tweets will receive a response, and how much discussion they will generate. We learned that discussions are often started by very polarised posts and by users with a large broadcast spectrum (e.g., large number of followers, appearing on many subscription lists). Time of posting also turned out to be a crucial feature in seeding discussions. Overall, user features rendered better performance than content features for identifying discussion seed posts and for predicting discussion activity levels. For this work we used the explicit reply chain of posts in Twitter, rather than the retweet chain as in [11] and [12].

Unlike any of the above works, in this paper we mix content and user features with user-focus features for measuring the concentration of users on particular topic threads. The features that influence the dynamics and evolution of communities that are backboned by social networks and relationships (e.g., Facebook, Twitter) could be different from those that have content as a core (e.g., Digg, Slashdot). To this end, in this paper we compare our findings from Boards.ie against Twitter.

III. Feature Engineering

Our approach for anticipating discussions in community forums functions using two stages: 1) identifying seed posts, and; 2) predicting the level of discussion that seed posts will generate. We define a *seed post* as a thread starter posted by a user in a given forum that yields at least one reply, where the replier is not the original seed author, while a *non-seed post* is a thread starter that yields no replies from others. For the first stage of our approach we seek to find features that are consistent with seed posts and enable their distinction from non-seeds. To this end we explore three distinct feature sets that can be leveraged to describe such posts, and enable the key differences in such features to be observed for seeds and non-seeds. Our feature sets cover user features, content features and focus features.

A. User Features

User features describe the author of the post, seeking to identify key behavioural attributes that are synonymous with seed and non-seed posts. We define 5 user features as follows:

- *In-degree:* For the author of each post (seed or non-seed), this feature measures the number of incoming connections to the user.
- *Out-degree:* This feature measures the number of outgoing connections from the user.
- *Post Count:* Measures the number of posts that the user has made over the previous 6-months.
- *User Age:* Measures the length of time that the user has been a member of the community;
- *Post Rate:* Measures the number of posts made by the user per day.

B. Content Features

Independent of user and focus features, content features describe solely the post itself, identifying attributes that the content of seed posts should contain in order to start a discussion. We define 7 content features as follows:

- *Post Length:* Number of words in the post.
- *Complexity:* Measures the cumulative entropy of terms within the post to gauge the concentration of language and its dispersion across different terms. Let n be the number of unique terms within the post p and f_i is the frequency of term t within p, therefore complexity is given by:

$$\frac{1}{n}\sum_{i=1}^{n} f_i(\log n - \log f_i) \quad (1)$$

- *Readability:* Gunning fog index using average sentence length (ASL) [13] and the percentage of complex words (PCW): $0.4 * (ASL + PCW)$ This feature gauges how hard the post is to parse by humans.
- *Referral Count:* Count of the number of hyperlinks within the post.
- *Time in day:* The number of minutes through the day that the post was made. This feature is used to identify key points within the day that are associated with seed or non-seed posts.
- *Informativeness:* The novelty of the post's terms with respect to other posts. We derive this measure using the Term Frequency-Inverse Document Frequency (TF-IDF) measure.
- *Polarity:* Assesses the average polarity of the post using Sentiwordnet.[2] Let n denote the number of unique terms in post p, the function $pos(t_.)$ returns the positive weight of the term $t_.$ from the lexicon and $neg(t_.)$ returns the negative weight of the term. We therefore define the polarity of p as:

$$\frac{1}{n}\sum_{i=1}^{n} pos(t_i) - neg(t_i) \tag{2}$$

C. Focus Features

Our third feature set measures the concentration of the post author in individual forums. Our intuition is that by gauging the focus of a user we will be able to capture his/her expertise and areas of interest. Posts made on Boards.ie are published in distinct forums, where each forum can be regarded as focussing on a specific topic. Therefore, one would expect the majority of users to concentrate and network in a few forums. Using the forum information in our dataset we explore the use of 2 features as follows:

- *Forum Entropy:* Measures the concentration of users across the forums that they posts in. A higher value indicates a more random distribution - indicating that the user participates in many different forums - while a lower value corresponds to the user being attached to a lower number of forums. Let F_{v_i} be all the forums that user v_i has posted in and $p(f_.|v_i)$ be the conditional probability of v_i posting in forum $f_.$ - we can derive this using the post distribution of the user - therefore we define the Forum Entropy (H_F) of a given user as:

$$H_F(v_i) = -\sum_{j=1}^{|F_{v_i}|} p(f_j|v_i)\log p(f_j|v_i) \tag{3}$$

  To contextualise the derivation of this feature, consider the example shown in Figure 1 where a user ($U1$) has made 3 posts ($P1$, $P2$ and $P3$) where each post is made within a single forum. Given a new post $P4$, which we wish to predict as either being a seed post or not, we can look at the past focus of the user to gauge the concentration of their content. Our intuition is that a larger forum entropy will correlate with non-seeds, given that the user does not focus their discussions in a few, select forums.
- *Forum Likelihood:* Assesses the likelihood that the user will post within a forum given the past forum distribution of the user. Unlike Forum Entropy, the Forum Likelihood combines both known information about the user with new incoming information in the form of a likelihood assumption. We use Laplace smoothing of the probability distribution to account for unseen information - i.e., for the user posting in a forum that they have never posted in before - thereby calculating the maximum likelihood estimate of the user (v_i) posting in a forum f_j. Let $c(f_p, v_i)$ be the number of times that user v_i has posted in the forum (f_p) of the new post (p), let P_{v_i} denote the set of posts by v_i and F be the set of all forums. Using the example from Figure 1, we are able to take into account the piece of information that the new post $P4$ is made within forum $F2$. We define the forum likelihood of forum f_p given v_i as:

$$P(f_p|v_i) = \frac{c(f_p, v_i) + 1}{|P_{v_i}| + |F|} \tag{4}$$

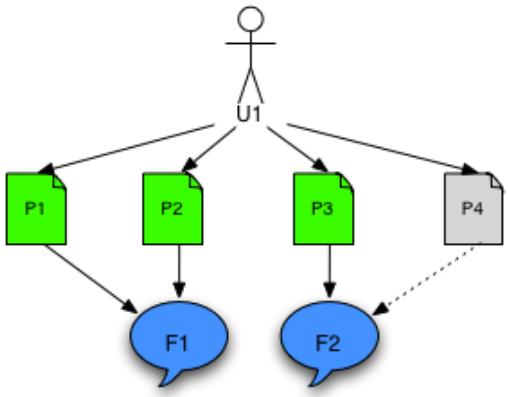


Fig. 1. Distribution of the user posts and the forums they appear in. P1, P2 and P3 are previous posts by the user in forums F1 and F2, while P4 is a new post.

IV. Dataset: Boards.ie

The dataset used for our experiments and analysis was kindly provided by Boards.ie, an Irish community message board that has been in existence since 1998. The message board covers a wide variety of topics and forums and is not limited to a single domain of interest. We were provided with a full collection of the data since the message board's inception through to 2008. However, due to the sheer scale and magnitude of this collection we decided to analyse data from a single year. For this year we chose 2006, given its high

[2]http://sentiwordnet.isti.cnr.it/

activity levels, and because it has already been used in other investigations (e.g., [14]).

Boards.ie does not provide explicit social relations between community members, unlike for example Facebook and Twitter. We followed the same strategy proposed in [3] for extracting social networks from Digg, and built the Boards.ie social network for users, weighting edges cumulatively by the number of replies between any two users.

TABLE I
DESCRIPTION OF THE BOARDS.IE DATASET

Posts	Seeds	Non-Seeds	Replies	Users
1,942,030	90,765	21,800	1,829,465	29,908

In order to take derive our features we required a window of n-days from which the social graph can be compiled and relevant measurements taken. Based on previous work over the same dataset in [14], we used a similar window of 188 days (roughly 6-months) prior to the post date of a given seed or non-seed post. For instance, if a seed post p is made at time t, then our window from which the features (i.e., user and focus features) are derived is from $t - 188$ to $t - 1$. In using this heuristic we ensure that the features compiled for each post are independent of future outcomes and will not bias our predictions - for example a user may increase their activity following the seed post which would not be a true indicator of their behaviour at the time the post was made. Table I summarises the dataset and the number of posts (seeds, non-seeds and replies) and users contained within.

V. CLASSIFICATION: DETECTING SEED POSTS

Predicting discussion activity levels are often hindered by including posts that yield no replies. We alleviate this problem by differentiating between seed posts and non-seeds through a binary classification task. Once seed posts have been identified we then attempt to predict the level of discussion that such posts will generate. To this end, we look for the best classifier for identifying seed and non-seed posts and then search for the features that played key roles in distinguishing seed posts from non-seeds, thereby observing key features that are associated with discussions.

A. Experimental Setup

For our experiments we are using the previously described dataset collected from Boards.ie containing both seeds and non-seeds throughout 2006. For our collection of posts we built the content, user, and focus features listed in section III from the past 6 months of data leading up to the date on which the post was published - thereby ensuring no bias from future events in our dataset. We split the dataset into 3 sets using a 70/20/10% random split, providing a training set, a validation set and a test set.

Our first task was to perform *model selection* by testing four different classifiers: SVM, Naive Bayes, Maximum Entropy and J48 decision tree, when trained on various individual feature sets and their combinations: user features, content features and focus features. This model selection phase was performed by training each classifier, together with the combination of features, using the 70% training split and labelling instances in the held out 20% validation split.

Once we had identified the best performing model - i.e., the classifier and combination of feature set that produces the highest F_1 value - our second task was to perform *feature assessment*, thereby identifying key features that contribute significantly to seed post prediction accuracy. For this we trained the best performing model from the model selection phase over the training split and tested its classification accuracy over the 10% test split, dropping individual features from the model and recording the reduction in accuracy following the omission of a given feature. Given that we are performing a binary classification task we use the standard performance measures for such a scenario: precision, recall and f-measure - setting $\beta = 1$ for an equal weighting of precision and recall. We also measure the area under the Receiver Operator Characteristic curve to gauge the relationship between recall and fallout - i.e., false negative rate.

TABLE II
RESULTS FROM THE CLASSIFICATION OF SEED POSTS USING VARYING FEATURE SETS AND CLASSIFICATION MODELS

		P	R	F_1	ROC
User	SVM	0.775	**0.810**	**0.774**	0.581
	Naive Bayes	0.691	0.767	0.719	0.540
	Max Ent	0.776	0.806	0.722	0.556
	J48	**0.778**	0.809	0.734	**0.582**
Content	SVM	0.739	0.804	0.729	0.511
	Naive Bayes	0.730	0.794	0.740	0.616
	Max Ent	0.758	0.806	0.730	**0.678**
	J48	**0.795**	**0.822**	**0.783**	0.617
Focus	SVM	0.649	0.805	0.719	0.500
	Naive Bayes	**0.710**	0.737	**0.722**	0.588
	Max Ent	0.649	0.805	0.719	**0.586**
	J48	0.649	**0.805**	0.719	0.500
User + Content	SVM	0.790	0.808	0.727	0.509
	Naive Bayes	0.712	0.772	0.732	0.593
	Max Ent	0.767	0.807	0.734	0.671
	J48	**0.795**	**0.821**	**0.779**	**0.675**
User + Focus	SVM	0.776	**0.810**	**0.776**	0.583
	Naive Bayes	0.699	0.778	0.724	0.585
	Max Ent	0.771	0.806	0.722	0.607
	J48	**0.777**	0.810	0.742	**0.617**
Content + Focus	SVM	0.750	0.805	0.729	0.511
	Naive Bayes	0.732	0.787	0.746	0.658
	Max Ent	0.762	0.807	0.731	**0.692**
	J48	**0.798**	**0.823**	**0.787**	0.662
All	SVM	0.791	0.808	0.727	0.510
	Naive Bayes	0.724	0.780	0.740	0.637
	Max Ent	0.768	0.808	0.733	0.688
	J48	**0.798**	**0.824**	**0.792**	**0.692**

B. Results: Model Selection

1) Model Selection with Individual Features: The results from our first experiments are shown in Table II. Looking first at individual feature sets - e.g., SVM together with user features - we see that content features yield improved predictive performance over user and focus features. On discussion forums *content appears to play a more central role*

in driving discussions rather than merely the networking or reputation of the person. Some of the features we analyse (e.g., Informativeness, Complexity, Referral Count - see section III) act as proxies of posts' quality or attractiveness.

SVM provides the best performance when using user features, indicating the utility of discriminative classification models in identifying key differences between seeds and non-seeds in terms of the characteristics of the post author. The results of F_1 levels suggest that using *focus features on their own are not sufficient to accurately distinguish seeds from non-seeds*. In terms of the area under the ROC curve, focus features are not worst, indicating that the *false positive rate is worse when using user features* - confirmed by the lower levels of precision leading to more non-seeds being labelled as seeds. This contrasts with our results from Twitter, where user features were found to be better than content features for predicting seed posts [2].

2) Model Selection with Combined Features: Next we focus on merging features together, and find that *combining user and content features improves classification over the sole use of user features* for J48 and Maximum Entropy, while reducing it for SVM and Naive Bayes. For J48, *performance also improves when merging user and focus features* in comparison to using these feature sets in isolation. Focus features allow users' posting history and their topical concentration to be taken into account. *Combining content features with focus features sees the largest improvements* over the use of solitary feature sets when using the J48 classifier, which significantly outperforms other classifiers trained using the same features - using the sign test with a significance setting of $p < 0.01$ and testing the null hypothesis that there was no difference between the results of the J48 classifier and other classifiers.

3) Model Selection with All Features: Finally, in terms of F_1 levels, we found that we achieve the *best performance when combining all features* together into the same classification model using J48 and Maximum Entropy. We tested the results from the J48 classifier trained using all features against the next best performing feature set combination - content and focus features - and found the improvement in the F_1 level to be statistically significant with $p < 0.01$. As a baseline measure, we computed the weighted average between two classifiers, one where all posts are classed as seeds and the other where all posts are classed as non-seeds. As each classifier returns a recall of 1 - given that it identifies all seeds and non-seeds respectively - we were interested in the precision level that such an approach yields. We found precision to be 0.686 thereby indicating the significant improvement that we gain, in precision, over such a baseline through the use of our classification approach with different feature sets - only dropping below this precision level when solely focus features are used with the J48 classifier.

C. Results: Feature Assessment

The goal of this second experiment is to assess the contribution of each feature individually when identifying seed posts, exploring our research question: *which features are key for stimulating discussions?* Based on our previous model selection phase we identified the J48 decision tree classifier together with all features as being the best performing model. We trained this model over the 70% training split and classified instances within the held-out test split of 10% of the dataset - thereby ensuring independence from our previous model selection phase. This model, that we shall refer to hereafter as Ψ, formed our baseline against which the effects of dropping a single feature from Ψ are compared (Figure 2).[3] *Dropping Forum Likelihood, Informativeness, Readability and Referral Count results in a clear reduction in F_1 and ROC levels*, whilst dropping Post Count, Post Rate and Forum Entropy sees no reduction in performance.

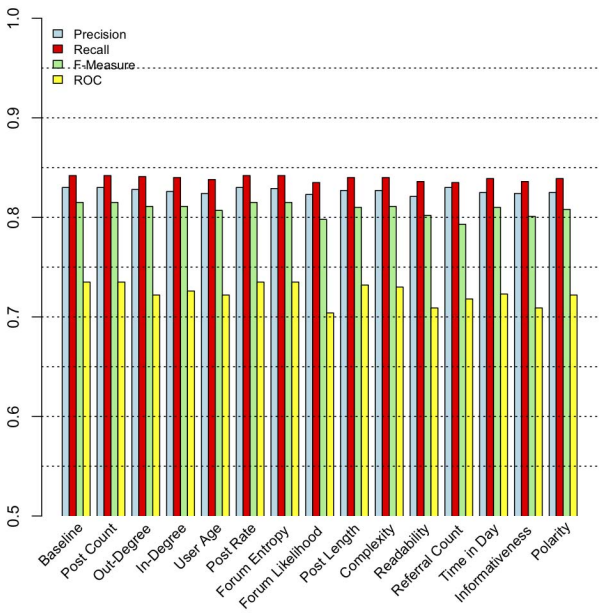


Fig. 2. Reduction in performance levels as individual features are dropped from the J48 classification model trained using all features

To test the significance of the reduction in F_1 levels we performed the sign test using the null hypothesis that there is no difference between Ψ and $\Psi_{feature}$, and rejecting this hypothesis if there is a significant difference. Table III show the F_1 values produced when individual features are removed from Ψ and the significance of those reductions.

To provide a more detailed insight into the relationship between statistically significant features and seed posts we plotted boxplots of the distribution of each feature with regards to both the seed posts (labelled '*pos*') and non-seeds (labelled '*neg*') within the training split - as shown in Figure 3 for the top-5 features that produce the greatest reduction in F_1 levels. *Non-seeds correlate with high referral counts*, indicating that eliciting a reply is more likely if the post did not contain many

[3]We shall hereafter denote the model with a given feature omitted using the convention $\Psi_{feature}$ for brevity.

TABLE III
REDUCTION IN F_1 LEVELS AS INDIVIDUAL FEATURES ARE DROPPED FROM THE J48 CLASSIFIER

Feature Dropped	F_1
-	0.815
Post Count	0.815
In-Degree	0.811*
Out-Degree	0.811*
User Age	0.807***
Post Rate	0.815
Forum Entropy	0.815
Forum Likelihood	0.798***
Post Length	0.810**
Complexity	0.811**
Readability	0.802***
Referral Count	0.793***
Time in Day	0.810**
Informativeness	0.801***
Polarity	0.808***

Signif. codes: p-value $<$ 0.001 *** 0.01 ** 0.05 * 0.1 .

hyperlinks (e.g., ads and spams). This contrasts with work in Twitter which found that tweets containing many links were more likely to get '*retweeted*' [11].

The boxplot for Forum Likelihood shows a correlation between seed posts and higher values of the likelihood measure, suggesting that *users who frequently post in the same forums are more likely to start a discussion*. Also, If a user often posts in discussion forums, while concentrating on only a few select forums, then the likelihood that a new post is within one of those forums is high.

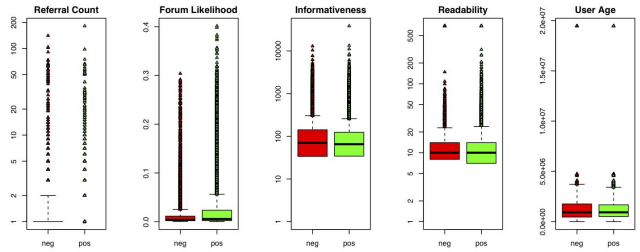


Fig. 3. Boxplots showing the correlation of feature values with seed and non-seed posts within the training split

VI. REGRESSION: PREDICTING DISCUSSION ACTIVITY

Early detection of lengthy discussions helps analysts and managers to focus attention to where activity and topical debates are about to occur. In this section we predict the level of discussion activity that seed posts will generate and what features are key indicators of lengthy discussions. We use regression models that induce a function describing the relationship between the level of discussion activity and our user, content and focus features. By learning such a function we can identify patterns in the data and correlations between our dependent variable and the range of predictor variables that we have.

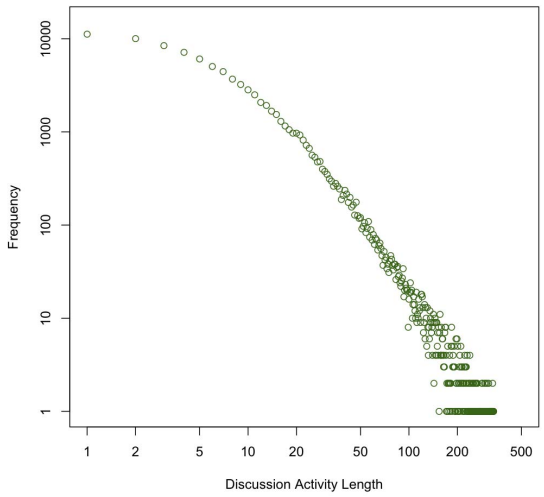


Fig. 4. Discussion Activity Length Distribution

A. Experimental Setup

Forecasting the exact number of replies (discussion activity) is limited if the distribution of known reply lengths has a large skew to either the minimum or maximum. For predicting popular tweets, Hang et al [12] adopted a multiclass classification setting to deal with the large skew in the dataset by predicting retweet count ranges. We have a similar scenario in our Boards.ie dataset, where a large number of seed posts yield fewer than 20 replies (Figure VI). In such cases utilising standard regression error measures such as Relative Absolute Error produces inaccurate assessments of the predictions due to using a simple predictor based on the mean of the target variables.

In our experiments we instead use the Normalised Discounted Cumulative Gain (nDCG) at varying rank positions, looking at the performance of our predictions over the top-k documents where $k = \{1, 5, 10, 20, 50, 100\}$. NDCG is derived by dividing the Discounted Cumulative Gain (DCG) of the predicted ranking by the actual rank defined by (iDCG). DCG is well suited to our setting, given that we wish to predict the most popular posts and then expand that selection to assess growing ranks, as the measure penalises elements in the ranking that appear lower down when in fact they should be higher up. We define DCG formally, based on the definition from [9], as:

$$DCG_k = \sum_{i=1}^{k} \frac{rel_i}{\log_2(1+i)} \tag{5}$$

For our experiments we first identify the best performing regression model before moving onto analysing the coefficients of that model and the patterns in the data that lead to increased discussion activity. For our *model selection* phase we test three regression models: Linear regression, Isotonic

(weighted least-squares) and Support Vector Regression, each trained on our various feature sets both in isolation and combined. The experiment was conducted by training the models on the seed posts in the 70% training split and then predicting the discussion activity level of the seed posts in the 20% validation split, where the seed posts in the latter split were identified using the previously described identification technique involving the J48 classifier trained on all features.

B. Results

Figure 5 shows the results obtained by each model and feature set combination for nDCG@k at varying levels of k. For predicting the top-ranked position - i.e., nDCG@1 - the use of solely user features in the Linear and SVR models outperforms other feature sets (e.g., achieving 0.89 for Linear regression with user features). This indicates that the use of solely *user features provides an effective method of predicting the post that will yield the highest number of replies*. Figure 5(a) indicates that over all nDCG levels the combination of content and focus features achieves the highest average value. There is also very little variation in the values across the different settings of k - characterised by achieving the lowest standard deviation of all tested models ($\sigma = 0.020$) - thus indicating that this combination of features when used within a *Linear Regression model provides the best method for estimating discussion volume across varying ranks*.

To summarise the results in Figure 5, Table IV presents the average results for each model and feature set combination across all tested nDCG@k levels. For single feature sets the results show that SVR achieves optimum performance using just user features. This outcome is similar to the classification experiment above where the use of support vectors for inducing the regression function allows key differences to be observed between seed post authors that correlate with discussion volume. *Combining content and focus features in the Linear Regression model achieves the best performance over other feature combinations*, including the use of all features.

TABLE IV
AVERAGED NDCG@K LEVELS FOR DIFFERENT REGRESSION MODELS AND FEATURE SETS

	Linear	Isotonic	SVR
User	0.646	0.412	**0.702**
Content	0.433	0.506	**0.512**
Focus	**0.587**	0.567	0.587
User + Content	0.547	0.411	**0.710**
User + Focus	0.660	0.630	**0.722**
Content + Focus	**0.756**	0.642	0.628
All	0.687	0.630	**0.711**
Average	0.617	0.543	**0.645**

1) Feature Contributions: We now wish to explore the correlation of the features in the Linear Regression model with discussion volume to answer the question of *how do these features influence discussion length?* To identify such correlations we took the Linear regression model that was induced from Content and Focus features and analysed the coefficients in the model, where for each of the predictor variables the t-test was performed to gauge the significance of its inclusion in the model. Table V shows the results from this analysis.

TABLE V
SUMMARY OF A LINEAR REGRESSION MODEL INDUCED FROM CONTENT AND FOCUS FEATURES, THEIR COEFFICIENTS AND T-TEST RESULTS

	Coefficient	Error	t-Value	P($x >$t)
Forum Entropy	-0.2441	0.0381	-6.406	1.50^{-10} ***
Forum Likelihood	60.0807	2.1865	27.478	$<2^{-16}$ ***
Content Length	0.0369	0.0060	6.186	6.19^{-10} ***
Complexity	2.4775	0.3056	8.106	5.29^{-16} ***
Readability	0.0024	5.747^{-4}	4.142	3.45^{-5} ***
Referral Count	-0.1236	0.0449	-2.754	0.0059 **
Time in Day	7.98^{-5}	2.635^{-4}	0.303	0.7620
Informativeness	-0.0093	1.6394^{-3}	-5.643	1.67e-08 ***
Polarity	-4.0863	0.6478	-6.308	2.83^{-10} ***

Signif. codes: p-value $<$ 0.001 *** 0.01 ** 0.05 * 0.1 . 1

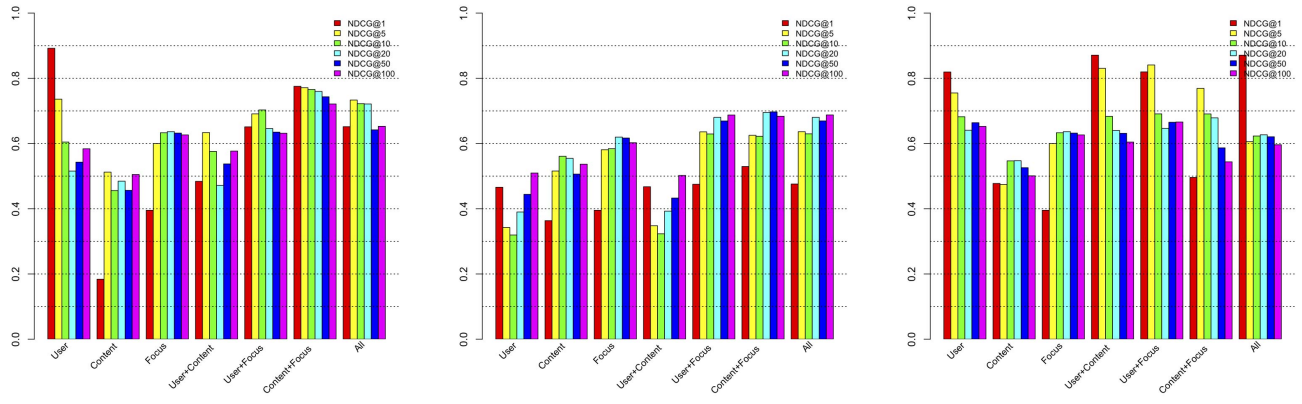
For focus features Table V shows that the greater the Forum Likelihood then the greater the discussion volume (significant at $p < 0.001$) - i.e., *if users focus their activity on a few forums, then their new posts in those forums are expected to generate a lengthy discussion*. For Forum Entropy the lower the entropy then the greater the discussion volume.

For the content features the results indicate that *the greater the complexity of posts, the greater the discussion volume*, suggesting that using more expressive language to stimulate discussions and a wider vocabulary leads to lengthier discussions. The negative coefficient for Referral Count indicates that greater discussion activity is correlated with fewer hyperlinks being shared within the post. This correlation could be attributed to spam content on discussion boards often containing many hyperlinks, a common occurrence on community discussion forums where users promote events and their own content.

VII. DISCUSSION AND FUTURE WORK

The features we selected in this study cover various internal parameters than can influence response and discussion levels in the community. External parameters, such as events in similar forums, admin or technical changes in the system, and media stories, could all have an impact on which topics will receive more attention from the community. In our work we assumed that posts followed the topic of the forum that contained them. However, some posts could be off-topic and as a result they can little attention. Tracking topics of posts in forums can help to measure this feature. Our future work will explore the use of topic models to describe the focus of a user given the topics that they discuss.

One of our findings indicated that having many URLs in posts can negatively impact discussion levels in Boards.ie. Although this might be due to the association of spam with URLs, it could also be due to such posts not being in the form of questions (e.g. links to articles) and hence did not require a direct response and discussion.



(a) Linear Regression (b) Isotonic Regression (c) SVR

Fig. 5. Normalised Discounted Cumulative Gain measures for different regression models at different values of k and feature sets

The Boards.ie social network was generated from users' interactions, and hence its characteristics and meaning are likely to be different from those in Twitter and the like. This could render the user features in Boards.ie and Twitter less comparable.

We showed that the influence of certain features can be totally different from one community to another. Such experiments need to be repeated with bigger datasets and across many more online communities of similar natures and goals to see if the same patterns emerge.

VIII. Conclusions

This paper presented a two-stage approach to anticipate discussions in community forums. We explored a range of features and how they impacted upon identifying seed posts and how they are correlated with discussion activity. In exploring the question *what features are key for stimulating discussions?* we found content features to be better indicators of seed posts than user features, which is the opposite of our findings from analysing discussion dynamics in Twitter [2]. We also showed that if users concentrate their activities in certain forums then when they publish a new post in one such forum they can expect a lengthier discussion.

By implementing three different regression models trained on our explored features we were able to effectively predict the ordering of posts by their discussion activity levels, thereby allowing analysts to track the lengthiest discussions that will garner the greatest influence. We found that a combination of content and focus features provided the most consistent means to predict discussion activity levels across various rank levels. Analysis of the Linear regression model induced using content and focus features explored *how do features influence discussion length?* Correlations in the model suggest that as Forum Likelihood increases and Forum Entropy decreases at a similar rate then discussion activity levels increase.

Acknowledgment

The work of the authors was supported by the EU-FP7 projects WeGov (grant no. 248512) and Robust (grant no. 257859). Also thanks for Boards.ie for making their data available to this project.

References

[1] J. Lovett, C. A. Doty, V. Sehgal, S. Vittal, and E. Murphy, "US web analytics forecast, 2008 to 2014," 2009. [Online]. Available: http://www.forrester.com/rb/Research/us_web_analytics_forecast/,_2008_to_2014/q/id/53629/t/2

[2] M. Rowe, S. Angeletou, and H. Alani, "Predicting discussions on the social semantic web," in *Extended Semantic Web Conference*, Heraklion, Crete, 2011.

[3] H. Rangwala and S. Jamali, "Defining a Coparticipation Network Using Comments on Digg," *IEEE Intelligent Systems*, vol. 25, no. 4, pp. 36–45, 2010.

[4] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Commun. ACM*, vol. 53, no. 8, pp. 80–88, 2010.

[5] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman, "Knowledge sharing and Yahoo Answers: Everyone knows something," in *Proc WWW Conf*, 2008.

[6] J. Bian, Y. Liu, D. Zhou, E. Agichtein, and H. Zha, "Learning to Recognize Reliable Users and Content in Social Media with Coupled Mutual Reinforcement," in *18th Int WWW Conf*, April 2009.

[7] V. Gómez, A. Kaltenbrunner, and V. López, "Statistical analysis of the social network and discussion threads in slashdot," in *WWW '08: Proceeding of the 17th international conference on World Wide Web*. New York, NY, USA: ACM, 2008, pp. 645–654. [Online]. Available: http://dx.doi.org/http://doi.acm.org/10.1145/1367497.1367585

[8] S. Siersdorfer, S. Chelaru, W. Nejdl, and J. San Pedro, "How useful are your comments?: analyzing and predicting youtube comments and comment ratings," in *19th Int WWW Conf*, NY, USA, 2010.

[9] C.-F. Hsu, E. Khabiri, and J. Caverlee, "Ranking Comments on the Social Web," in *Int Conf Computational Science and Engineering (CSE) 2009*, 2009.

[10] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring User Influence in Twitter: The Million Follower Fallacy," in *4th Int AAAI Conf on Weblogs and Social Media*, 2010.

[11] B. Suh, L. Hong, P. Pirolli, and E. H. Chi, "Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network," in *Proc IEEE Second Int Conf on Social Computing (SocialCom)*, 2010.

[12] L. Hong, O. Dan, and B. D. Davison, "Predicting popular messages in twitter," in *Proc 20th Int WWW Conf*, New York, NY, USA, 2011.

[13] R. Gunning, *The Technique of Clear Writing*. McGraw-Hill, 1952.

[14] J. Chan, C. Hayes, and E. Daly, "Decomposing Discussion Forums using Common User Roles," in *Proc We Science Conf (WebSci)*, 2010.